Wikimpacts 1.0: A new global climate impact database based on automated information extraction from Wikipedia (in prepare)

Ni Li[1,2], Wim Thiery[1], Shorouq Zahra[3,6], Mariana Madruga de Brito[4], Koffi Worou[5,6], Murathan Kurfali[3,6], Seppe Lampe[1], Paul Munoz[1], Clare Flynn[5,6], Camila Trigoso[1], Joakim Nivre[3,6,7], Jakob Zscheischler[2,8,9], and Gabriele Messori[5,6,10]

Research background

Background:
- Bias in the current impact dataset such as EM-DAT
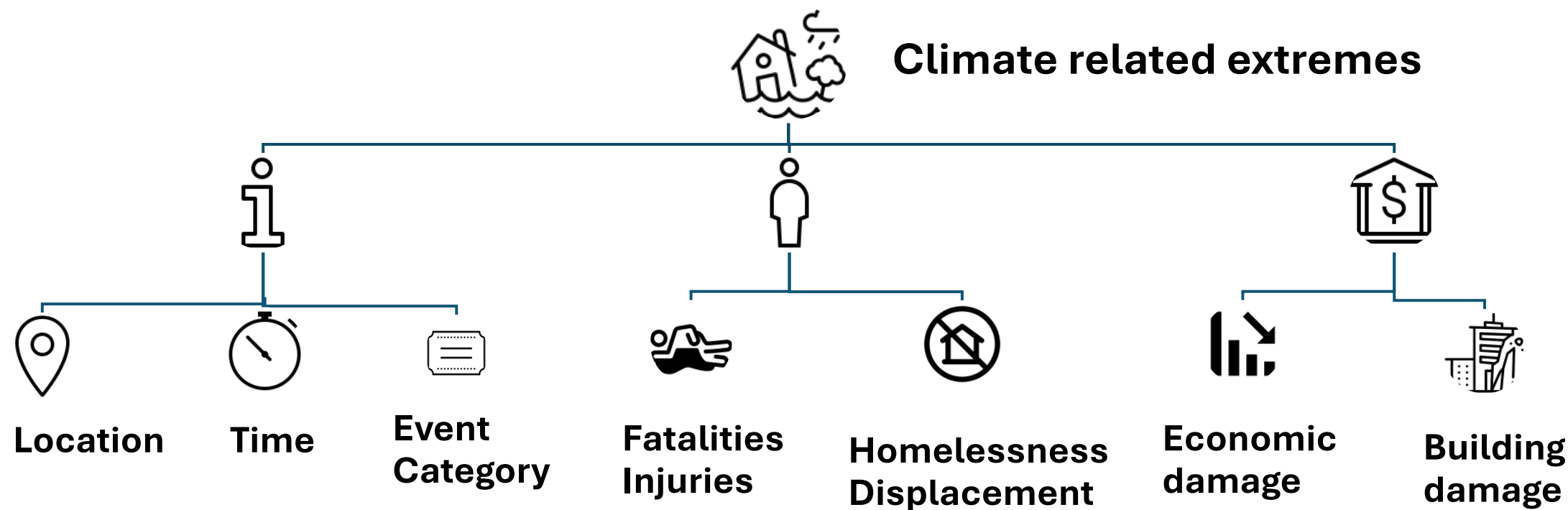- Increasing intensity and frequency of some extreme events (IPCC, 2023)

Wikimpacts:
- Mining Wikipedia with natural language processing and machine learning techniques
- Date, location, and socioeconomic impacts
- Uncertainty assessment

Table: Features of current disaster impact datasets

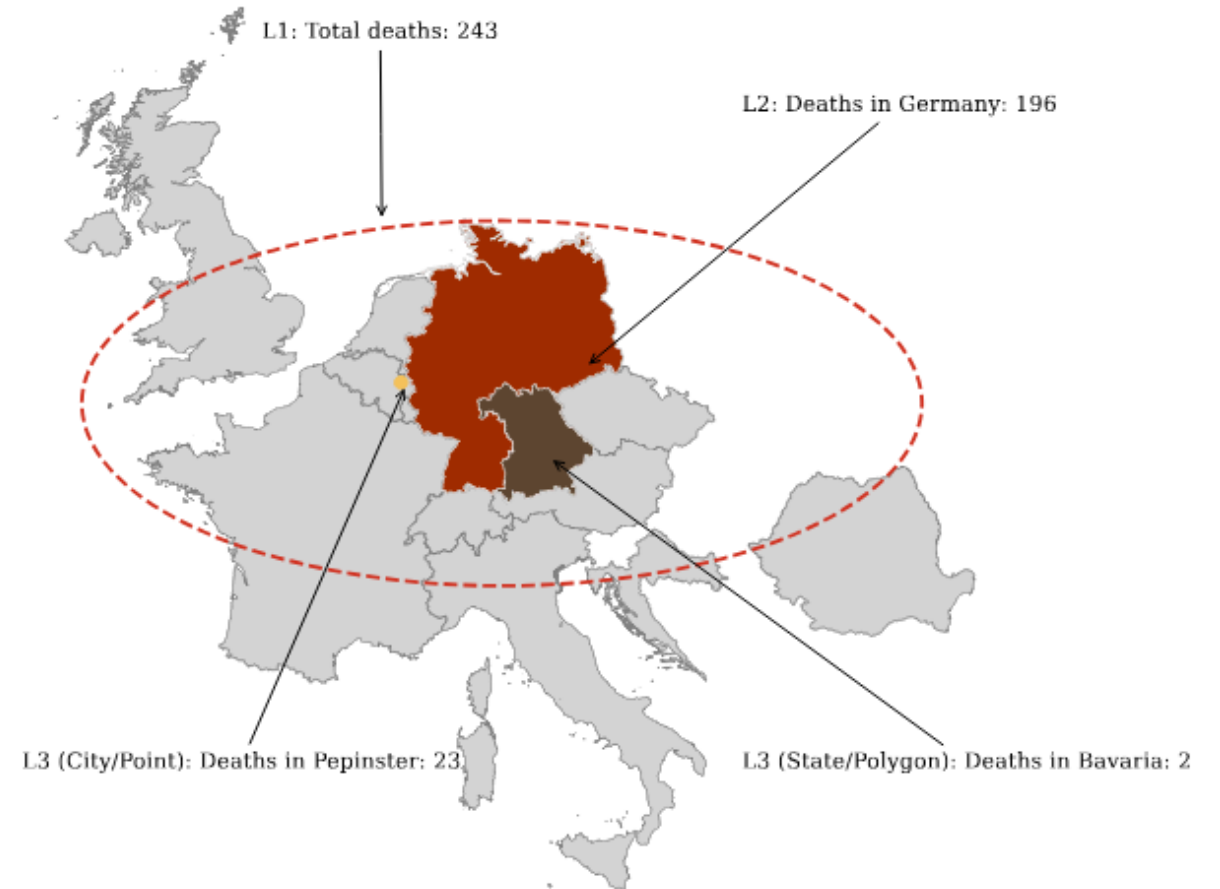| Dataset	Temporal coverage and data resource	Criteria to record	Ownership and Accessibility
EM-DAT	1900-present, UN agencies, IFRC, World Bank, reinsurers, press, news agencies, etc.	10 or more people killed or injured; 100 people affected; declaration of a state of emergency; or a call for international assistance	Centre for Research on the Epidemiology of Disasters, Belgium Free
NatCatSERVICE	79 AD-present, Property claims service, insurance clients, UN agencies, World Bank, press, academia, etc	The occurrence of human injury (loss of life, injury, homelessness) or property damage	Munich Re, Germany, Private
Sigma	1970-present, Property claims service, insurance clients, UN agencies, World Bank, press, academia, etc.	For the 2016 reporting year - insured losses: 19.9 million USD for maritime disasters, 39.8 million USD for aviation, 49.5 million USD for other losses, or economic losses: 99 million USD or Casualties: 20 dead or missing, 50 injured, 2000 homeless	Swiss Re Group, Swissland, Private
GLIDE	1930-present, UN agencies, IFRC, World Bank, reinsurers, press, news agencies, etc.	over 10 fatalities, over 100 affected, declaration of the state of emergency, or call for international assistance	Asian Disaster Reduction Center, Japan, Private

Wikimpacts 1.0 database content

Climate related extremes

Location

Time

Event Category

Fatalities Injuries

Homelessness Displacement

Economic damage

Building damage

Wikimpacts 1.0

- 2928 climate extreme events from English Wikipedia (flood, drought, storm, wildfire, tornado, extreme temperature)

- 3-level impact information
  - Event level, national level and sub-national level

- Automated information extraction pipeline (GPT4o model)

- Deaths, injuries, homeless, displaced, affected, buildings damaged, insured damage and total damage

2021 European Floods L1, L2 and L3 deaths overview in Wikimpacts 1.0 database

L1: Total deaths: 243

L2: Deaths in Germany: 196

L3 (City/Point): Deaths in Pepinster: 23

L3 (State/Polygon): Deaths in Bavaria: 2

Workflow

1. Document selection
- Keywords filter
- BERT text classifer
- Manual checking
- Single-event article selection

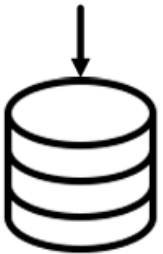Flood, Storm, Wildfire, Drought, etc
5046 articles are related to climate disasters
3368 articles are single-event articles

2. Information extraction (GPT4o)
- Prompt 1
- Prompt 2
  ...
- Prompt N

2 Prompts for Basic information, main event, hazard, countries and time
8 Prompts for Impact information: deaths, injuries, economic damage etc

3. Post-processing
- Data conversion
- Text normalization
- Geoparsing

Numerical data normalization: two->2, hundreds of -> 200-900
Date and Location normalization

4. Database consolidation
- Consistency checking
- Missing data
- Currency conversion
- Inflation adjustment

L1 >= L2 >=L3
USD|EUR
Adjusted to 2024 inflation

Evaluation
156 events in test set, 97 with L2/L3 annotated, score 0-1

Table 7. Results of the L2 evaluation on the test set, for each field within the impact categories. The Weighted Score represents the average across all fields in a given impact category, which are given an equal weighting. For example, the Weighted Score "0.4221" for "Deaths" is the mean error of all the fields in this category.

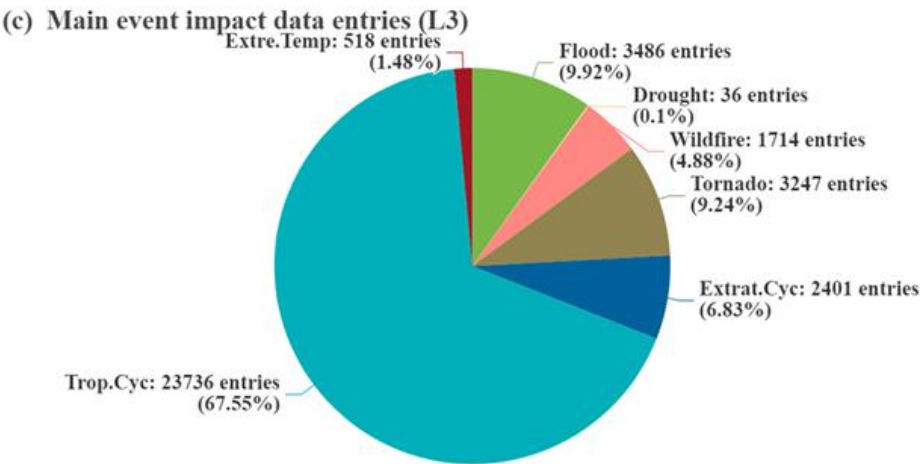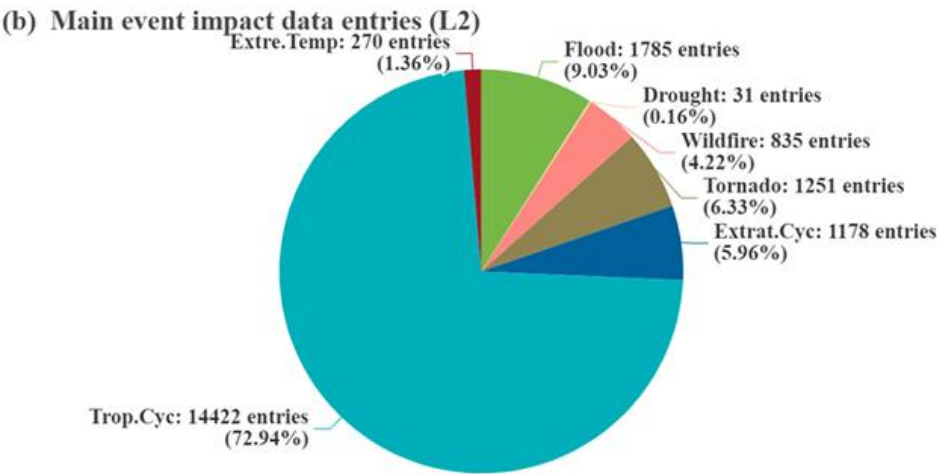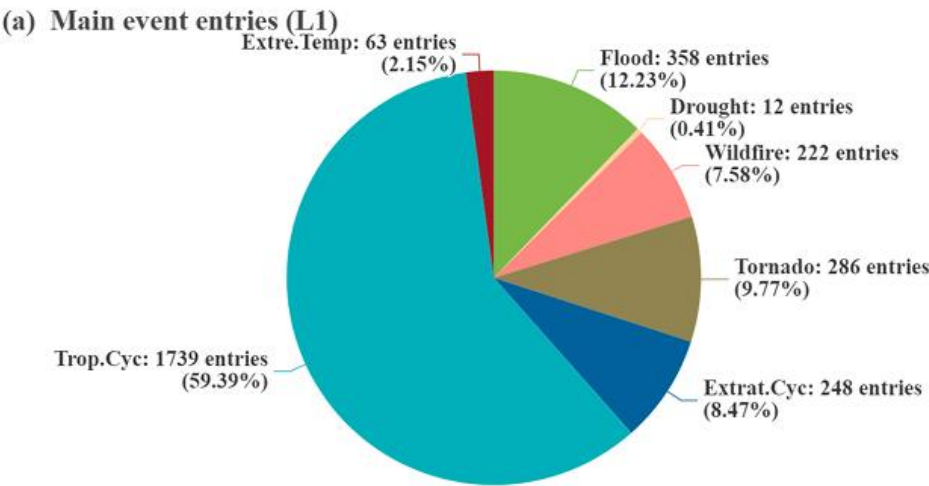|	Deaths	Injuries	Homeless	Displaced	Affected	Buildings_Damaged	Insured_Damage	Damage
Weighted_Score	0.4221	0.3171	0.3928	0.3709	0.4695	0.5125	0.5308	0.4772
Start_Date_Day	0.2310	0.2479	0.3051	0.3321	0.4247	0.4860	0.8210	0.5542
Start_Date_Month	0.2310	0.2479	0.3051	0.3321	0.4219	0.4832	0.8198	0.5550
Start_Date_Year	0.5446	0.3277	0.3517	0.3931	0.4665	0.5902	0.8198	0.6906
End_Date_Day	0.2244	0.1849	0.2288	0.1870	0.2987	0.4039	0.7477	0.4830
End_Date_Month	0.2244	0.1849	0.2288	0.1870	0.2946	0.4012	0.7477	0.4868
End_Date_Year	0.2508	0.2101	0.2331	0.1985	0.3163	0.4037	0.7523	0.4981
Administrative_Areas_Norm	0.7665	0.9076	0.9576	0.9351	0.9617	0.9358	0.9775	0.7887
Num_Min	0.6618	0.2724	0.4619	0.3864	0.5208	0.4543	0.1570	0.4030
Num_Max	0.6643	0.2702	0.4633	0.3865	0.5208	0.4541	0.1573	0.4023
Num_Unit	NA	NA	NA	NA	NA	NA	0.1757	0.4151
Num_Inflation_Adjusted	NA	NA	NA	NA	NA	NA	0.1892	0.4415
Num_Inflation_Adjusted_Year	NA	NA	NA	NA	NA	NA	0.0045	0.0075

Table 8. The L3 evaluation results of the test set, presented following the same format as in Table 7.

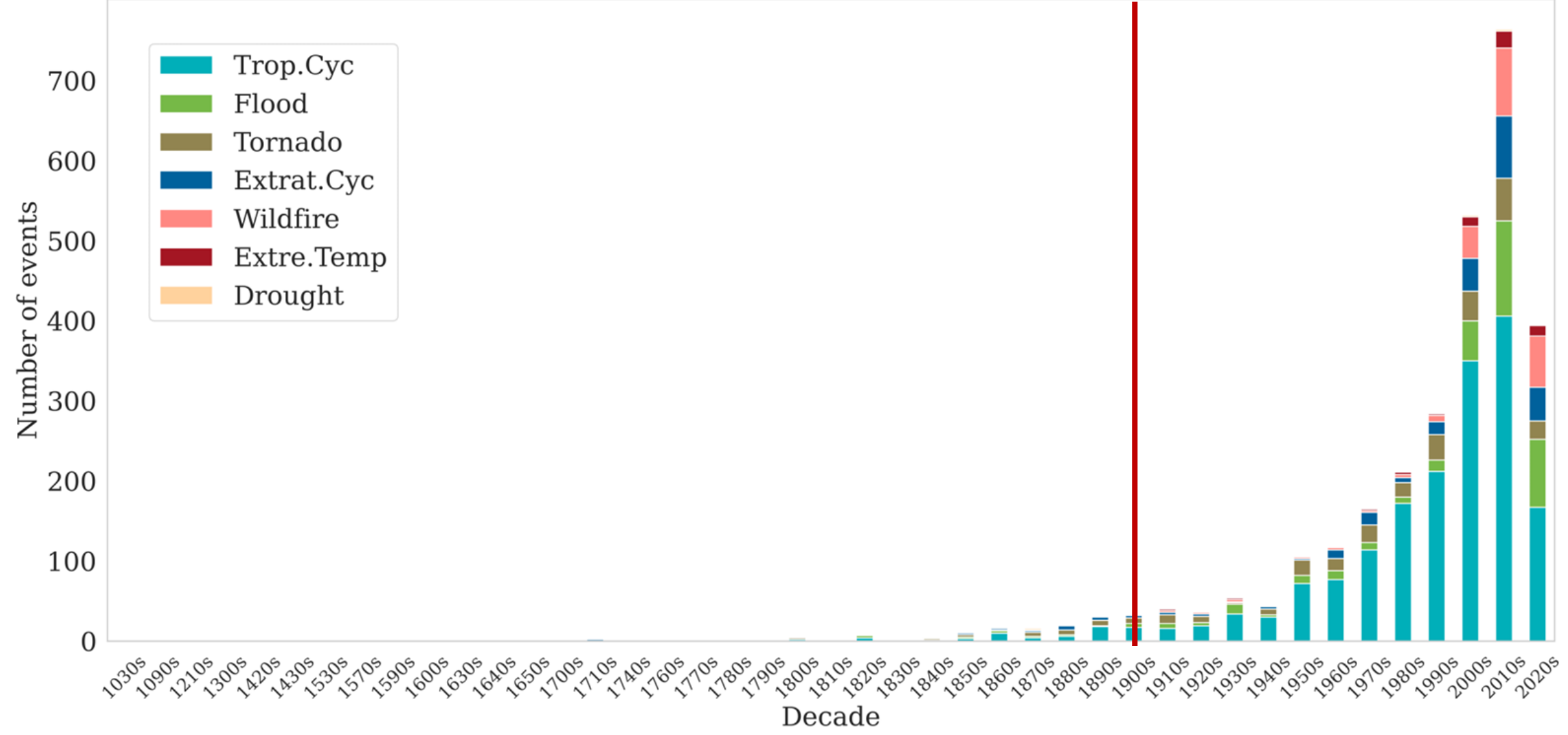|	Deaths	Injuries	Homeless	Displaced	Affected	Buildings_Damaged	Insured_Damage	Damage
Weighted_Score	0.4896	0.4228	0.4582	0.4684	0.5022	0.6031	0.5788	0.5371
Start_Date_Day	0.3222	0.2486	0.2699	0.3190	0.3081	0.5314	0.8110	0.6202
Start_Date_Month	0.3194	0.2486	0.2752	0.3178	0.3100	0.5326	0.8110	0.6260
Start_Date_Year	0.5916	0.3728	0.2888	0.4289	0.3161	0.6591	0.8171	0.6279
End_Date_Day	0.2728	0.1909	0.1962	0.1907	0.2207	0.4558	0.7530	0.5465
End_Date_Month	0.2743	0.1908	0.1962	0.1929	0.2204	0.4552	0.7530	0.5543
End_Date_Year	0.2915	0.1965	0.1962	0.2031	0.2249	0.4692	0.7561	0.5562
Administrative_Area_Norm	0.7225	0.9162	0.9646	0.8858	0.9757	0.8629	0.9939	0.9806
Locations_Norm	0.8528	0.9552	0.9714	0.9387	0.9886	0.9194	0.9939	1.0000
Num_Min	0.6245	0.4541	0.6117	0.6056	0.7288	0.5734	0.1925	0.3450
Num_Max	0.6247	0.4546	0.6115	0.6016	0.7289	0.5721	0.1921	0.3450
Num_Unit	NA	NA	NA	NA	NA	NA	0.2134	0.3837
Num_Inflation_Adjusted	NA	NA	NA	NA	NA	NA	0.2317	0.3876
Num_Inflation_Adjusted_Year	NA	NA	NA	NA	NA	NA	0.0061	0.0097

Table 6. The L1 evaluation results on the gold standard test set. The Weighted Score represents the average across all fields, given an equal weighting. For instance, the score of 0.0256 in the "Main_Event" field corresponds to the average score for all 156 test set events within this category. Numbers closer to 0 indicate a close match between two entries, while numbers closer to 1 indicate a poorer match.

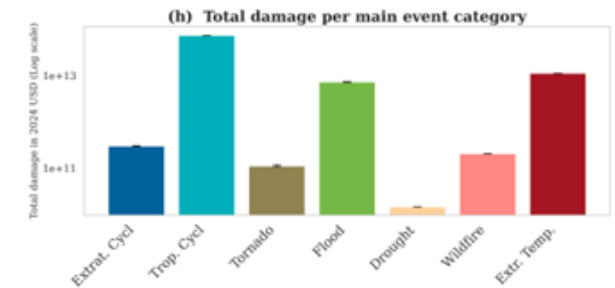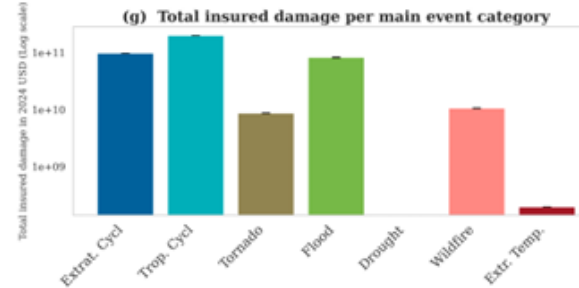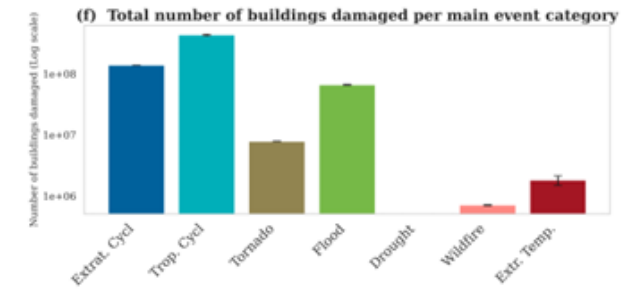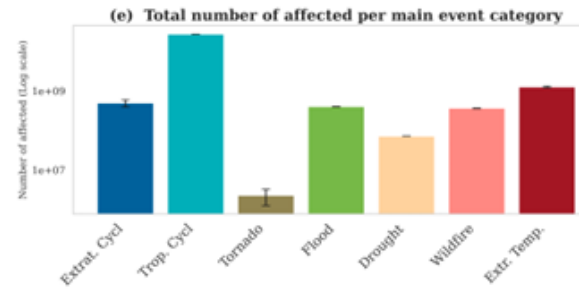| Field	Score
Weighted_Score	0.1431
Main_Event	0.0256
Hazards	0.2004
Start_Date_Day	0.0299
Start_Date_Month	0.0115
Start_Date_Year	0.0003
End_Date_Day	0.0463
End_Date_Month	0.0194
End_Date_Year	0.0066
Administrative_Areas_Norm	0.4843
Total_Deaths_Min	0.0374
Total_Deaths_Max	0.0236
Total_Injuries_Max	0.2118
Total_Injuries_Min	0.2115
Total_Homeless_Min	0.2559
Total_Homeless_Max	0.2528
Total_Displaced_Min	0.2950
Total_Displaced_Max	0.2963
Total_Affected_Min	0.3110
Total_Affected_Max	0.2993
Total_Buildings_Damaged_Min	0.2827
Total_Buildings_Damaged_Max	0.2797
Total_Insured_Damage_Min	0.1218
Total_Insured_Damage_Max	0.1218
Total_Insured_Damage_Unit	0.1474
Total_Insured_Damage_Inflation_Adjusted	0.1667
Total_Insured_Damage_Inflation_Adjusted_Year	0.0064
Total_Damage_Min	0.0706
Total_Damage_Max	0.0729
Total_Damage_Unit	0.0321
Total_Damage_Inflation_Adjusted	0.1026
Total_Damage_Inflation_Adjusted_Year	0.0128

Wikimpacts 1.0

(a) Main event entries (L1)

Extre.Temp: 63 entries (2.15%)
Flood: 358 entries (12.23%)
Drought: 12 entries (0.41%)
Wildfire: 222 entries (7.58%)
Tornado: 286 entries (9.77%)
Extrat.Cyc: 248 entries (8.47%)
Trop.Cyc: 1739 entries (59.39%)

(b) Main event impact data entries (L2)

Extre.Temp: 270 entries (1.36%)
Flood: 1785 entries (9.03%)
Drought: 31 entries (0.16%)
Wildfire: 835 entries (4.22%)
Tornado: 1251 entries (6.33%)
Extrat.Cyc: 1178 entries (5.96%)
Trop.Cyc: 14422 entries (72.94%)

(c) Main event impact data entries (L3)

Extre.Temp: 518 entries (1.48%)
Flood: 3486 entries (9.92%)
Drought: 36 entries (0.1%)
Wildfire: 1714 entries (4.88%)
Tornado: 3247 entries (9.24%)
Extrat.Cyc: 2401 entries (6.83%)
Trop.Cyc: 23736 entries (67.55%)
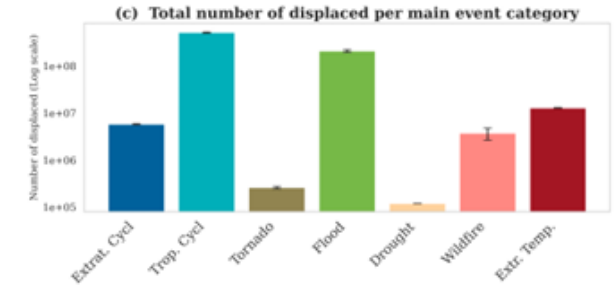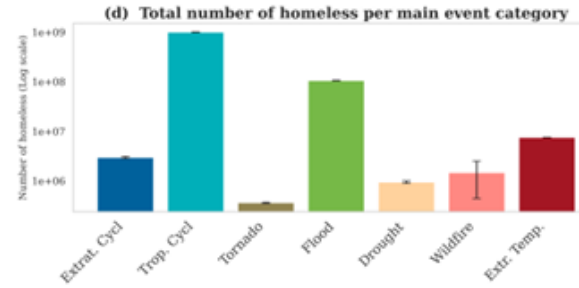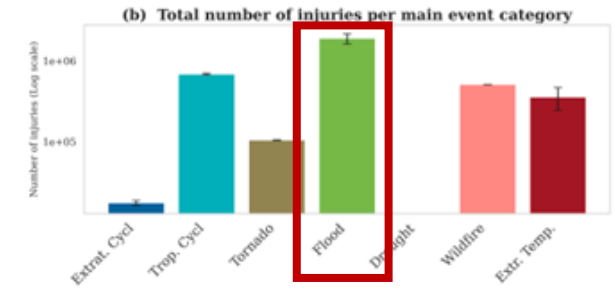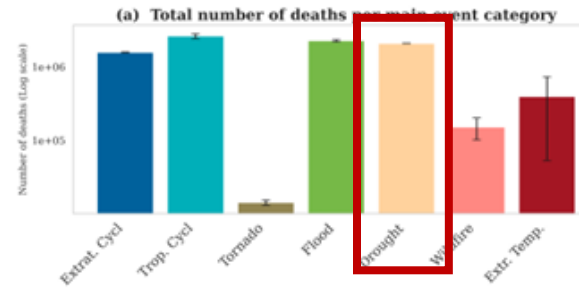
Wikimpacts 1.0

(a) Decadal distribution of main events (L1)

Wikimpacts 1.0

- Droughts led significant deaths

- Floods led most injuries

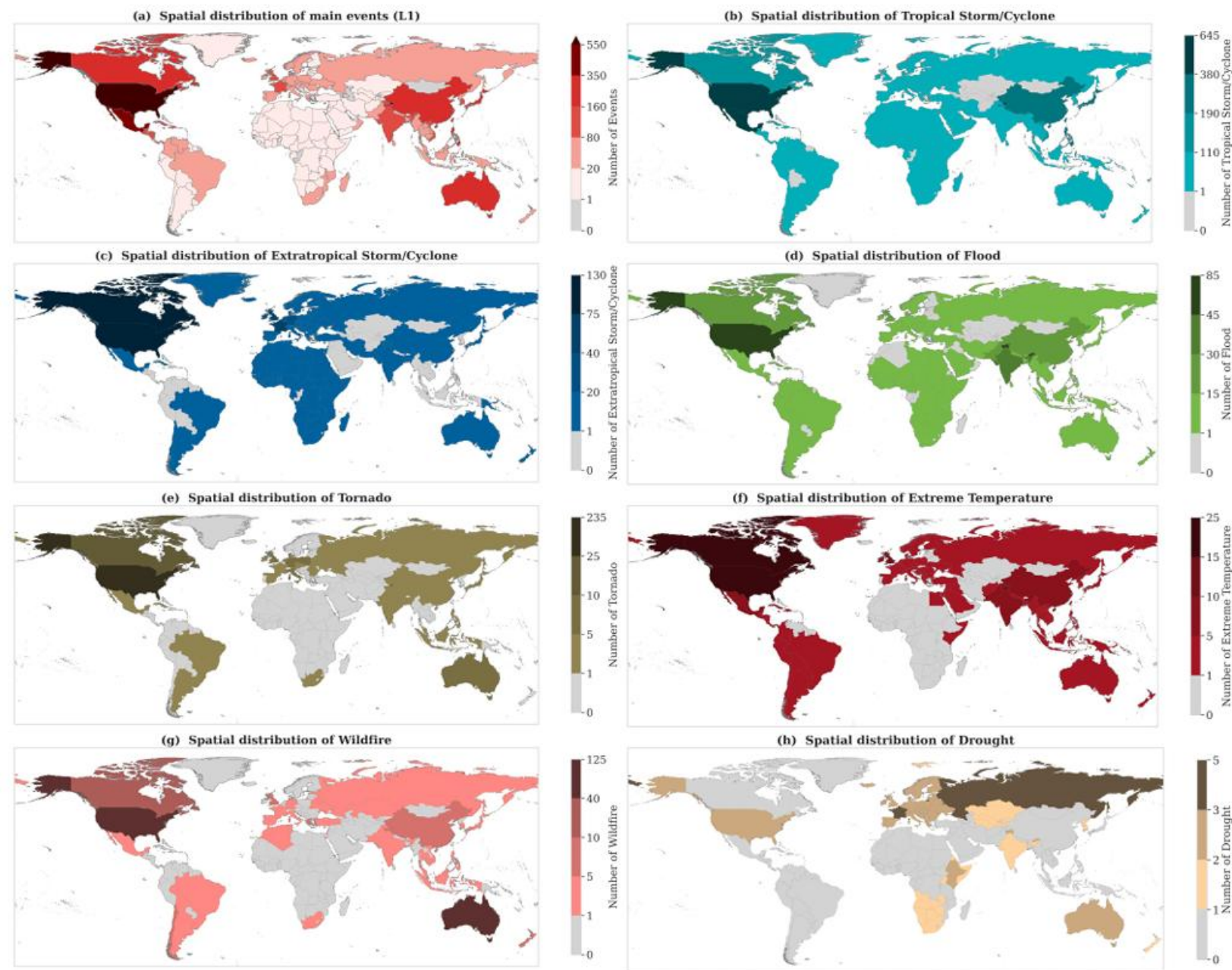- Tropical storms have the most severe impacts

(a) Total number of deaths per main event category

(b) Total number of injuries per main event category

(d) Total number of homeless per main event category

(c) Total number of displaced per main event category

(e) Total number of affected per main event category

(f) Total number of buildings damaged per main event category

(g) Total insured damage per main event category

(h) Total damage per main event category

Wikimpacts 1.0

L1 2928 events

Strong bias in US

Limited events in Africa

(a) Spatial distribution of main events (L1)

(b) Spatial distribution of Tropical Storm/Cyclone

(c) Spatial distribution of Extratropical Storm/Cyclone

(d) Spatial distribution of Flood

(e) Spatial distribution of Tornado

(f) Spatial distribution of Extreme Temperature

(g) Spatial distribution of Wildfire

(h) Spatial distribution of Drought

Wikimpacts 1.0

L2: 20,186 entries at the national level

L3: 36,394 entries at the sub-national level.

(a) Spatial distribution of impact data entries (L2)
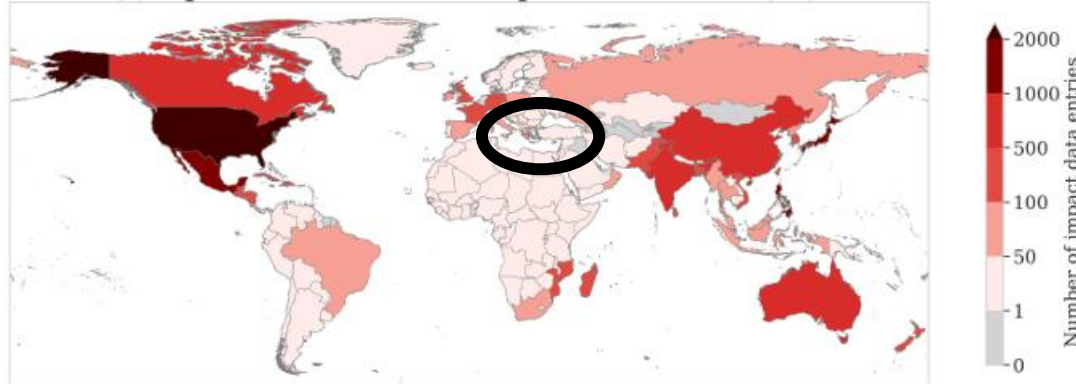
(b) Spatial distribution of impact data entries (L3 regions)

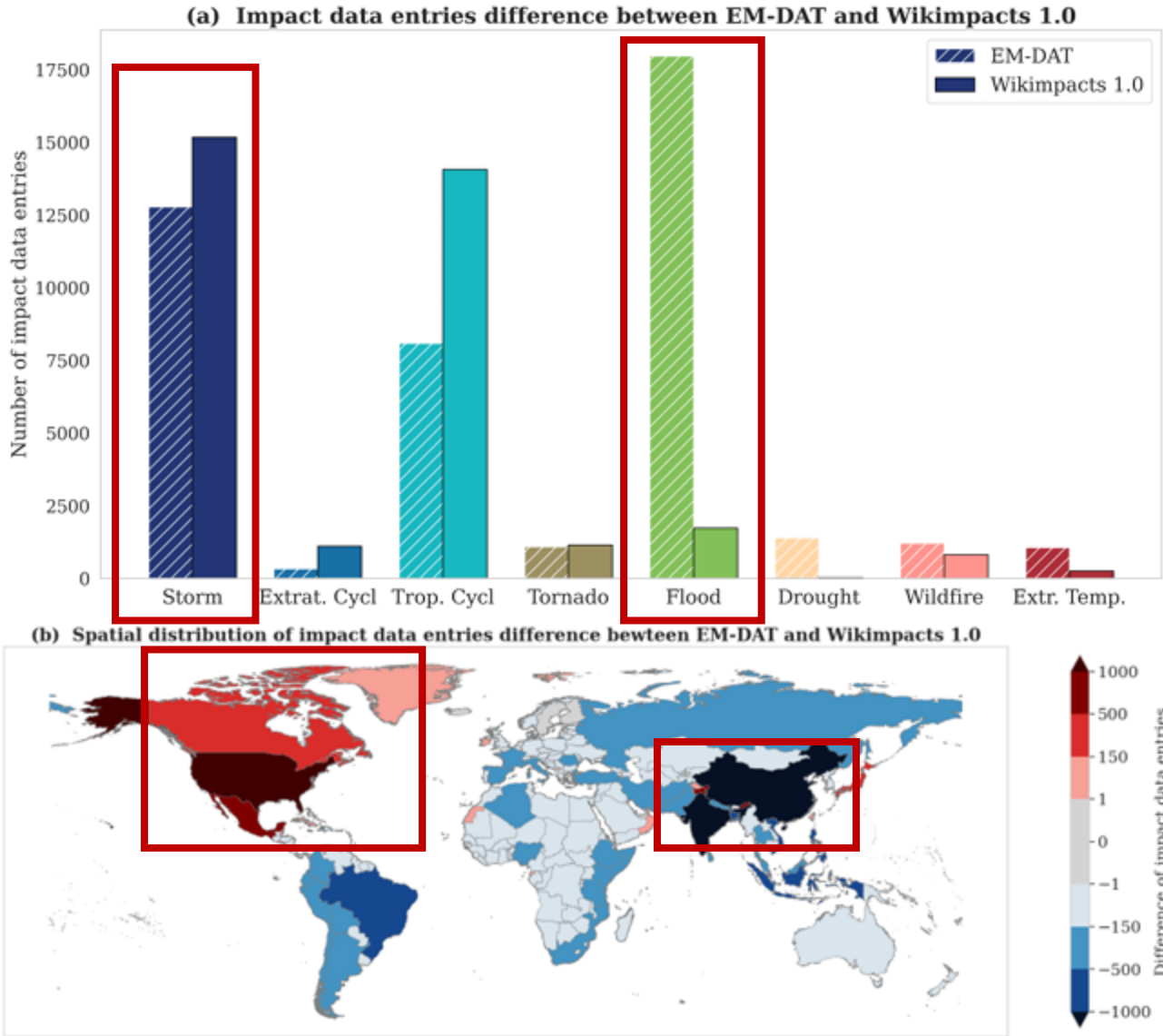(c) Spatial distribution of impact data entries (L3 cities)

Wikimpacts 1.0 L2 VS EM-DAT

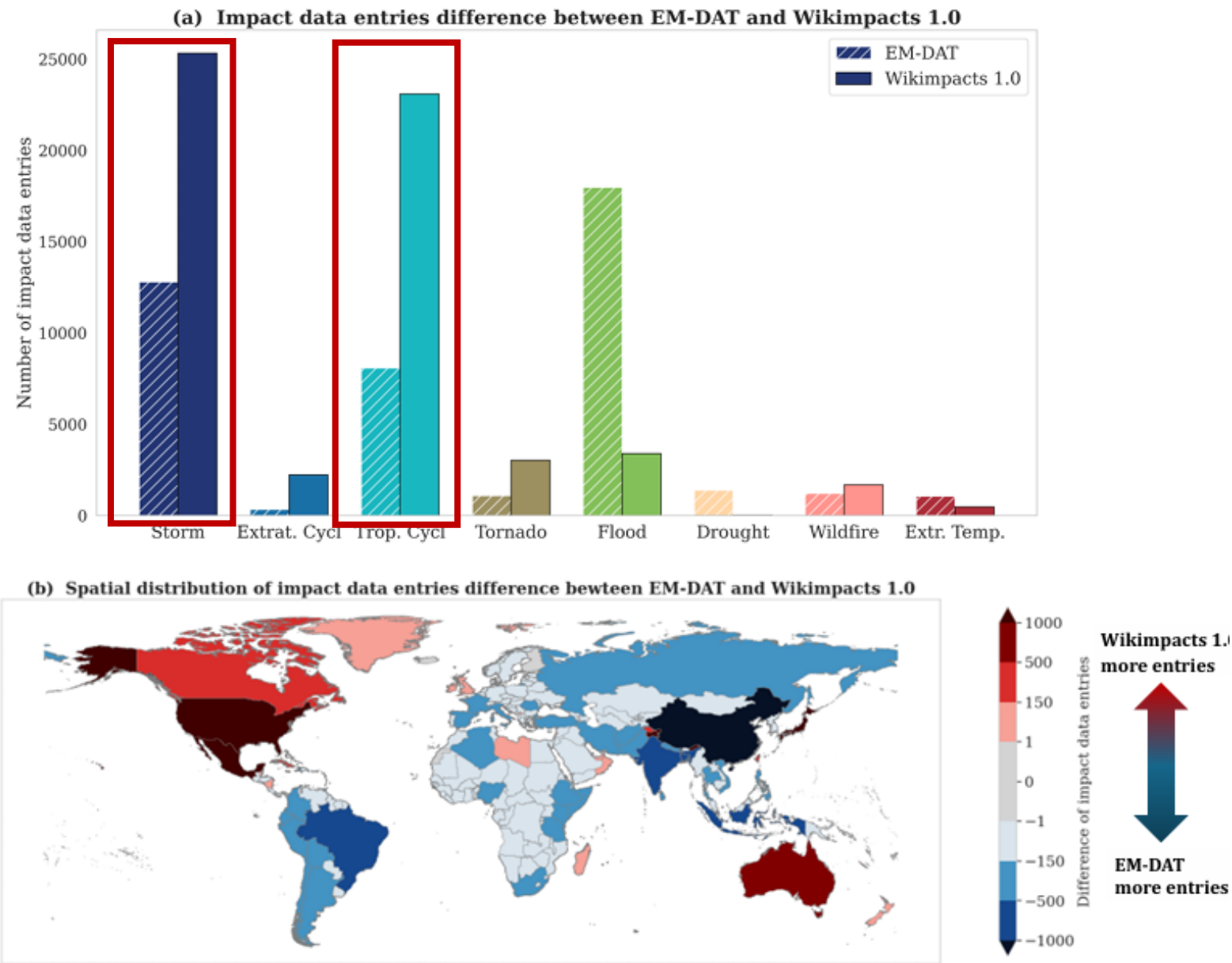More storm impact data

Less flood impact data

More impact data entries in North America

Less impact data in China and India

(a) Impact data entries difference between EM-DAT and Wikimpacts 1.0

(b) Spatial distribution of impact data entries difference bewteen EM-DAT and Wikimpacts 1.0

Wikimpacts L3 VS EM-DAT

More tropical cyclone sub-national impact data entries in Wikimpacts 1,0

(a) Impact data entries difference between EM-DAT and Wikimpacts 1.0

(b) Spatial distribution of impact data entries difference bewteen EM-DAT and Wikimpacts 1.0

Wikimpacts 1.0 VS EM-DAT –Impact Value Comparison
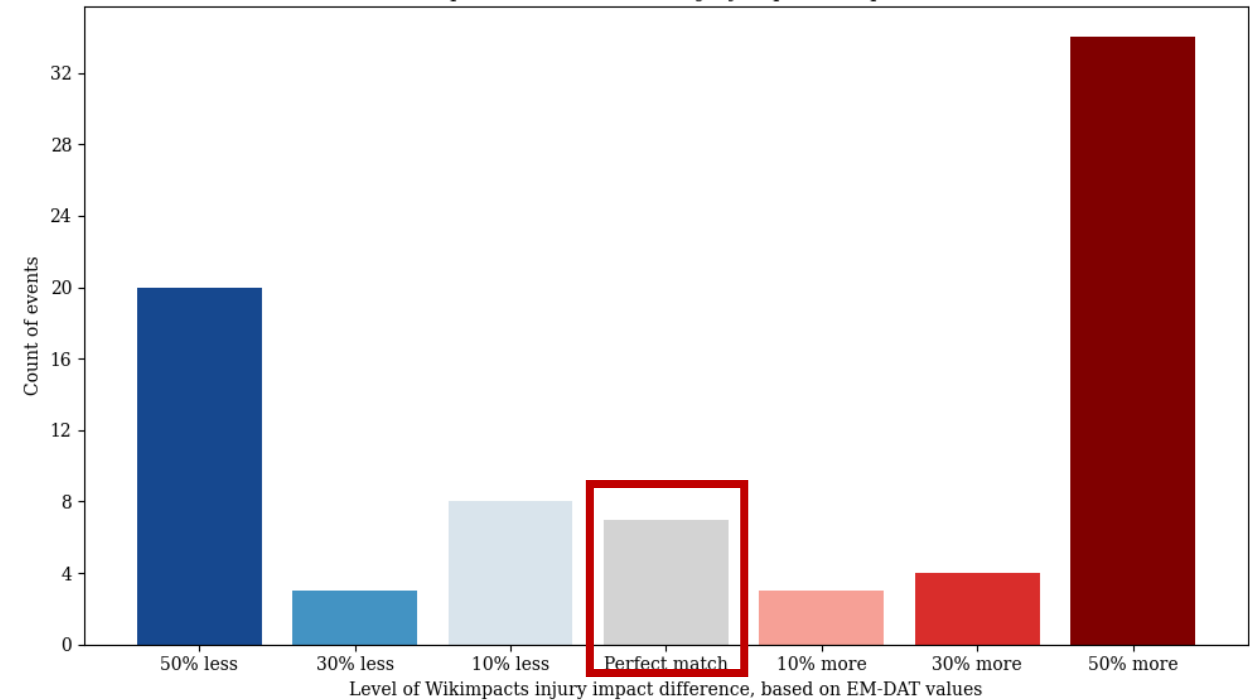
- Matching criteria: Event type, ISO code, start year, end year, start month and end month

  - 37 event death impact records are perfectly matched
  - 7 event injury impact records are perfectly matched

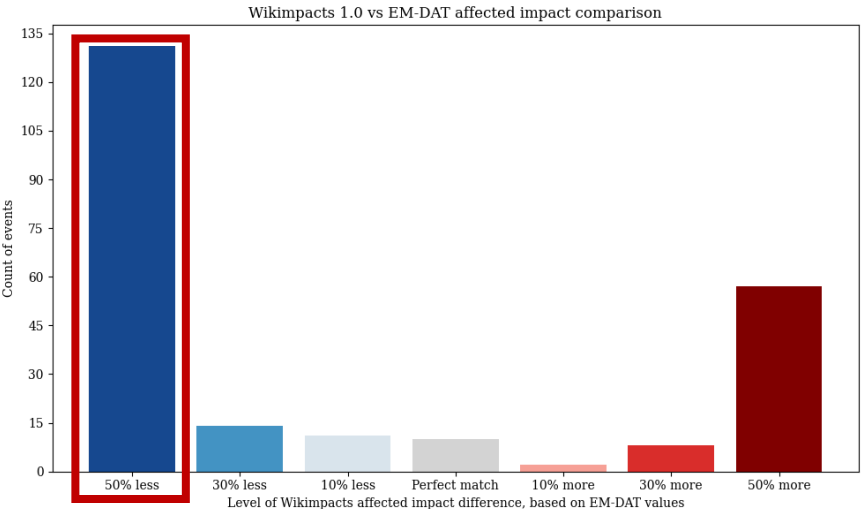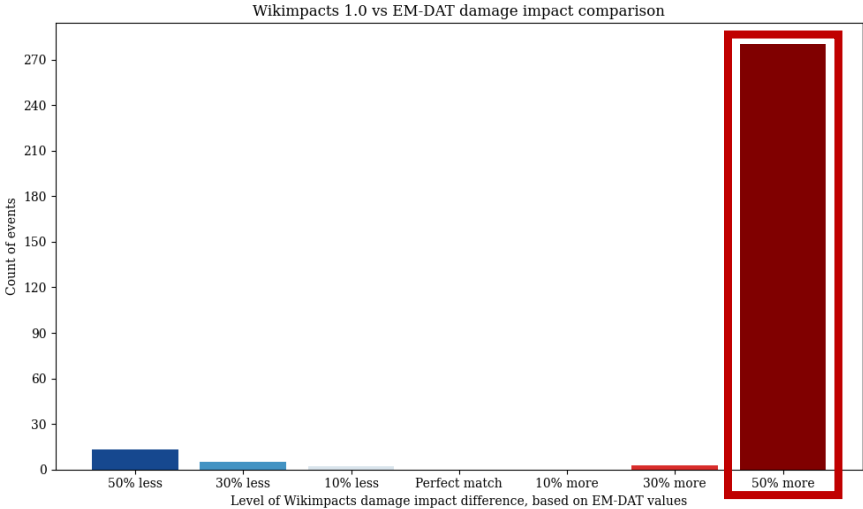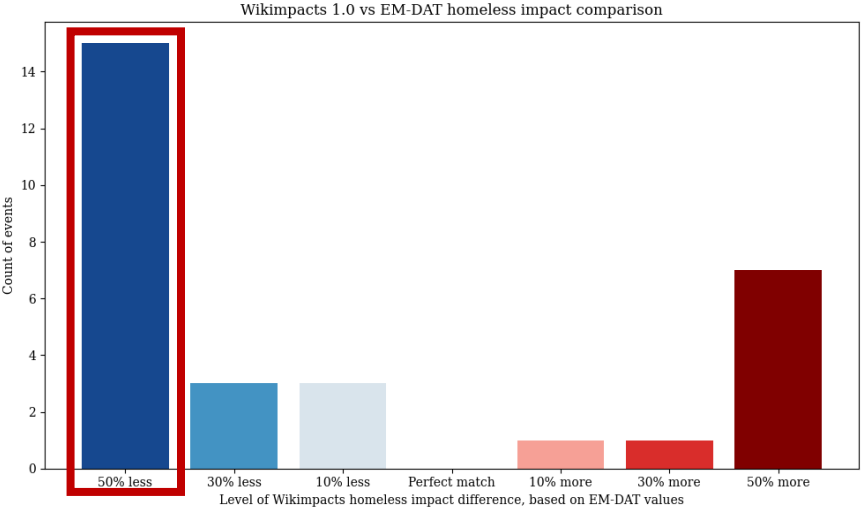Wikimpacts 1.0 vs EM-DAT death impact comparison
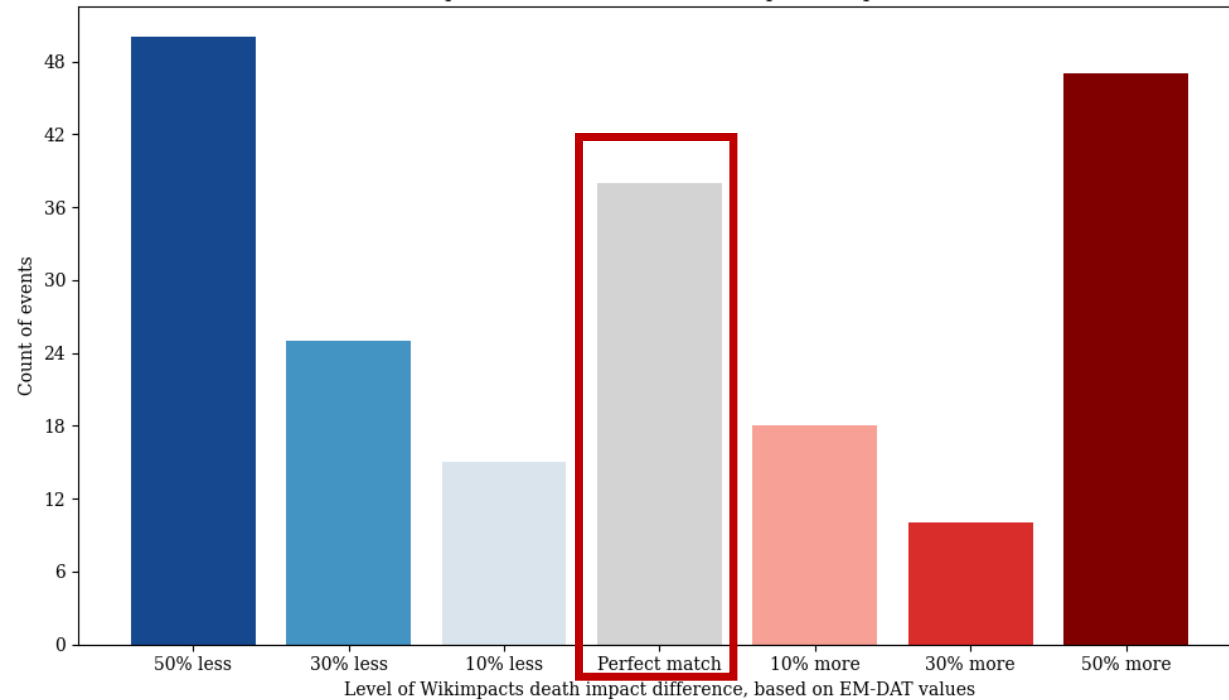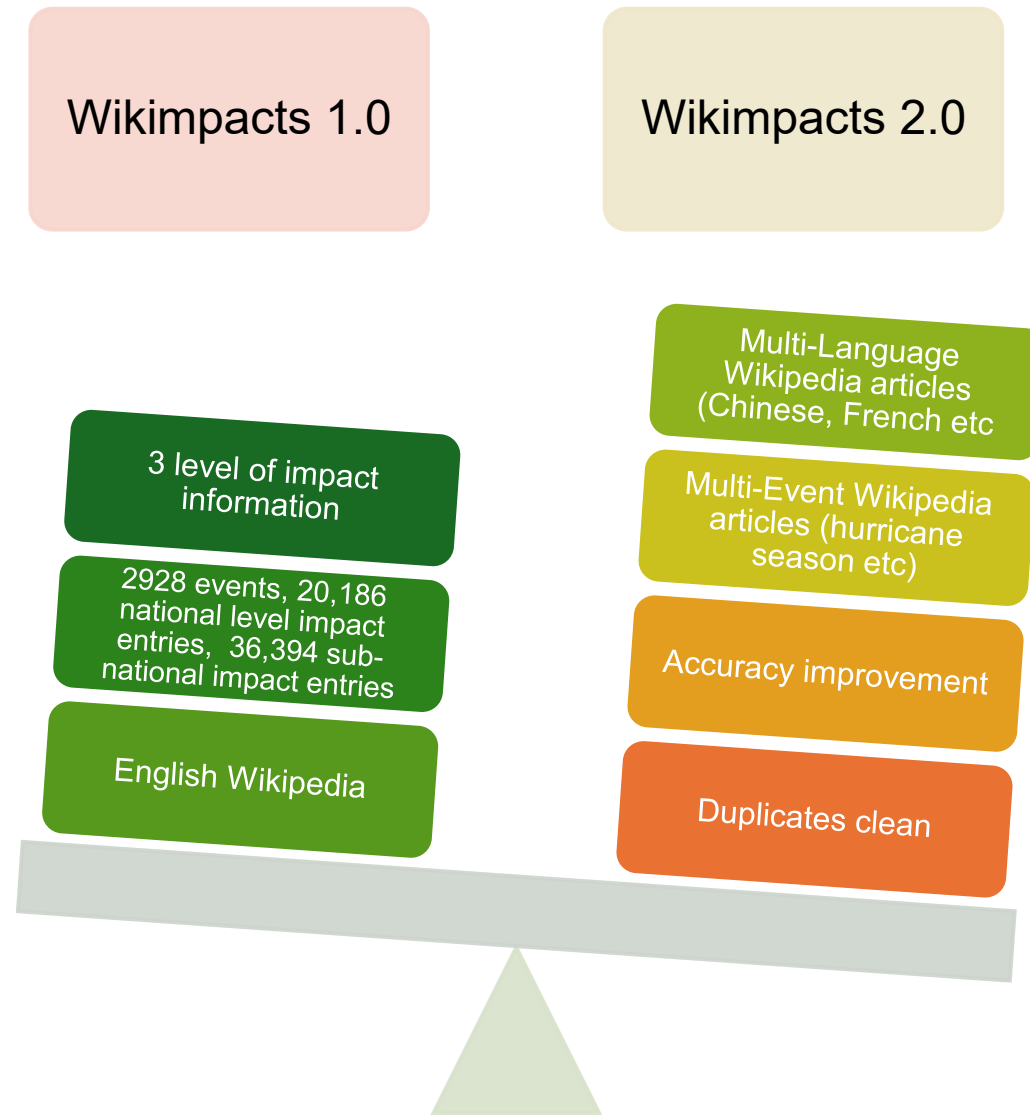
Wikimpacts 1.0 vs EM-DAT injury impact comparison

Wikimpacts 1.0 VS EM-DAT – Impact Value Comparison

- Homeless and affected people
  - Wikimpacts reports lower numbers

- Economic damage
  - Wikimpacts reports higher numbers

Wikimpacts 1.0 vs EM-DAT homeless impact comparison

Wikimpacts 1.0 vs EM-DAT damage impact comparison

Wikimpacts 1.0 vs EM-DAT affected impact comparison

Wikimpacts 1.0 → Wikimpacts 2.0

Wikimpacts 1.0

Wikimpacts 2.0

3 level of impact information

2928 events, 20,186 national level impact entries, 36,394 sub-national impact entries

English Wikipedia

Multi-Language Wikipedia articles (Chinese, French etc

Multi-Event Wikipedia articles (hurricane season etc)

Accuracy improvement

Duplicates clean

Contact us:
ni.li@vub.be